CLAIMS

What is claimed is:

1   1.   A method for evaluating and outputting a final clustering solution for

2   a plurality of multi-dimensional data records, said data records having multiple,

3   heterogeneous feature spaces represented by feature vectors, said method

4   comprising:

5   defining a distortion between two feature vectors as a weighted sum of

6   distortion measures on components of said feature vector;

7   clustering said multi-dimensional data records into k-clusters using a

8   "convex programming" formulation; and

9   selecting feature weights of said feature vectors.

1   2.   The method according to claim 1, wherein said selecting of feature

2   weights are optimized by an "objective" function to produce said solution of a

3   final clustering that simultaneously minimizes average intra-cluster dispersion and

4   maximizes average inter-cluster dispersion along all said feature spaces.

1   3.   The method according to claim 1, wherein said clustering includes

2   initially applying a local minima of said clustering.

1   4.   The method of claim 1, wherein said clustering comprises a k-means

2   clustering algorithm.

1   5.   The method of claim 2, wherein said minimizing distortion of individual

2   clusters includes taking said data records and iteratively determining *Voronoi*

3   partitions until said "objective" function, between two successive iterations, is

4   less than a specified threshold.

1   6.   The method of claim 1, wherein said clustering comprises analyzing word

2   data, and said feature vectors comprise multiple-word frequencies of said data

3   records.

1   7.   The method of claim 1, wherein said clustering comprises analyzing data

2   records having numerical and categorical attributes, and said feature vectors

3   comprise linearly-scaled numerical attributes and each q-ary categorical feature

4   using a 1-in-q representation of said data records.

1   8.   A method for evaluating and outputting a clustering solution for a plurality

2   of multi-dimensional data records, said data records having multiple,

3   heterogeneous feature spaces represented by feature vectors, said method

4   comprising:

5   defining a distortion between two said feature vectors as a weighted sum

6   of distortion measures on components of said feature vector;

7   clustering said multi-dimensional data records into k-clusters using a

8   "convex programming" formulation of a generalized k-means clustering function;

9   and

10   selecting optimal feature weights of said feature vectors by an "objective"

11   function to produce said solution of a final clustering that simultaneously

12   minimizes average intra-cluster dispersion and maximizes average inter-cluster

13   dispersion along all said feature spaces.

1   9.   The method of claim 8, wherein said clustering includes initially applying

2   a local minima of said clustering.

1   10.   The method of claim 8, wherein said minimizing distortion of individual

2   clusters includes taking said data records and iteratively determining *Voronoi*

3   partitions until said "objective" function, between two successive iterations, is

4   less than a specified threshold.

1   11.   The method of claim 8, wherein said clustering comprises analyzing word

2   data, and said feature vectors comprise multiple-word frequencies of said data

3   records.

1   12.   The method of claim 8, wherein said clustering comprises analyzing data

2   records having numerical and categorical attributes, and said feature vectors

3   comprise linearly-scaled numerical attributes and each q-ary categorical feature

4   using a 1-in-q representation of said data records.

1   13.   A computer system for data mining and outputting a final clustering

2   solution, wherein said system includes a memory for storing a database having a

3   plurality of multi-dimensional data records, each having multiple, heterogeneous

4   feature spaces represented by feature vectors, said system including a processor

5   for executing instructions comprising:

6   defining a distortion between two feature vectors as a weighted sum of

7   distortion measures on components of said feature vector;

8   clustering said multi-dimensional data records into k-clusters using a

9   "convex programming" formulation; and

10   selecting feature weights of said feature vectors.

1   14.   The system of claim 13, wherein said instruction for selecting of said

2   feature weights are optimized by implementing an "objective" function to produce

3   said solution of a final clustering that simultaneously minimizes average

4   intra-cluster dispersion and maximizes average inter-cluster dispersion along all

5   said feature spaces.

1   15.   The system of claim 13, wherein said instruction of said clustering

2   includes an instruction for initially applying a local minima of said clustering.

1   16.   The system of claim 13, wherein said instruction for clustering

2   includes instructions for implementing a k-means clustering algorithm.

1   17.   The system of claim 14, wherein said instruction for minimizing

2   distortion of individual clusters includes taking said data records and iteratively

3   determining *Voronoi* partitions until said "objective" function, between two

4   successive iterations, is less than a specified threshold.

1   18.   The system of claim 13, wherein said instruction for clustering includes

2   instructions for analyzing word data.

1   19.   The system of claim 13, wherein said instruction for clustering includes

2   instructions for analyzing data records having numerical and categorical attributes.

1   20.   A program storage device readable by machine, tangibly embodying a

2   program of instructions executable by said machine to perform a method for

3   evaluating and outputting a final clustering solution from a set of data records

4   having multiple, heterogeneous feature spaces represented as feature vectors, said

5   method comprising:

ARC9-2000-0078   34

6   defining a distortion between two feature vectors as a weighted sum of

7   distortion measures on components of said feature vector;

8   clustering said multi-dimensional data records into k-clusters using a

9   "convex programming" formulation; and

10   selecting feature weights of said feature vectors.

1   21.   The device of claim 20, wherein said selecting of feature weights are

2   optimized by an "objective" function to produce said solution of a final clustering

3   that simultaneously minimizes average intra-cluster dispersion and maximizes

4   average inter-cluster dispersion along all said feature spaces.

1   22.   The device of claim 20, wherein said clustering includes initially

2   applying a local minima of said clustering.

1   23.   The device of claim 20, wherein said clustering comprises a k-means

2   clustering algorithm.

1   24.   The device of claim 21, wherein said minimizing distortion of

2   individual clusters includes taking said data records and iteratively determining

3   *Voronoi* partitions until said "objective" function, between two successive

4   iterations, is less than a specified threshold.

1   25.   The device of claim 20, wherein said clustering comprises analyzing

2   word data, and said feature vectors comprise multiple-word frequencies of said

3   data records.

1   26.   The device of claim 20, wherein said clustering comprises analyzing

2   data records having numerical and categorical attributes, and said feature vectors

3   comprise linearly-scaled numerical attributes and each q-ary categorical feature

4   using a 1-in-q representation of said data records.